Disclaimer

The attached document is a preliminary contractor work product submitted by Tetra-Tech, Inc., to EPA's Office of Wastewater Management under Contract EP-C-05-046, Work Assignment 1-29. It presents preliminary technical recommendations for EPA to consider while developing its response to peer review comments on the draft *National Pollutant Discharge Elimination System Test of Significant Toxicity Implementation Document* (TST). While EPA considered the preliminary recommendations presented in this document, the Agency does not consider it to be a complete assessment of, or response to, comments received, and does not necessarily endorse the recommendations it contains. For this reason the attached document should not be interpreted as indicative of EPA's consideration of peer review comments.

EPA's final TST document reflects not only its consideration of Tetra-Tech's draft preliminary recommendations but also extensive input from experts and managers at EPA Headquarters, including the Office of Research and Development. Although it did not develop a separate Response to Comments document, EPA considered all of the peer review comments, and the final TST document reflects the Agency's consideration of those comments.

The final 2010 TST document is available at
http://www.epa.gov/npdes/pubs/wet_final_tst_implementation2010.pdf.

MEMORANDUM

Tetra Tech, Inc.
400 Red Brook Blvd., Suite 200
Owings Mills, MD 21117-6102
phone 410-356-8993
fax 410-356-9005

DATE: January 26, 2012

TO: Laura Phillips, EPA Work Assignment Manager

FROM: Jerry Diamond Ph.D., Tetra Tech Work Assignment Leader

SUBJECT: Response to external peer review comments from 2009

We are re-providing the attached draft Tetra Tech contract deliverable from January 23, 2009 entitled "Responses to Peer Review Comments: Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Approaches for Acute and Chronic Whole Effluent Toxicity". This Tetra tech deliverable was submitted under Work Assignment *1-29* (EPA OWM contract *EP-C-05-046*) as a draft document for EPA's internal review and consideration. Please note that the attached draft document does not include the many EPA decisions and changes made subsequent to external peer review comments, which are reflected in EPA's final TST document released in June 2010.

01/23/09

Responses to Peer Review Comments:

Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Approaches for Acute and Chronic Whole Effluent Toxicity

Submitted to:

Laura Phillips
USEPA
OWM
Washington, DC

Submitted by:

Tetra Tech, Inc.
400 Red Brook Blvd., Suite 200
Owings Mills, MD 21117

January 23, 2009

Table of Contents

EPA conducted a peer review of the "Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity" to obtain a critical, technical appraisal of the Test of Significant Toxicity (TST) Approach. From the eight reviewers approved, five were randomly selected and contacted to confirm that they could meet the time constraints of the project. Reviewers were charged with review of the document with reference to:

1) Document's Merit
2) Document's Responsiveness
3) Document's Data Analysis Basis
4) Document Conclusions
5) Document Overall Quality
6) Other Recommendations

Responses to comments from these five reviewers have been compiled into this peer review comment response document and are presented below.

Responses to General Comments

A summary of the results of the peer reviewer comments are as follows:

EPA Question 1 - Document Merit

Evaluate the conceptual soundness of the draft TST document's recommendations and the data analysis on which it is based. Is the draft TST approach an improvement over the current accepted hypothesis testing approach used in the NPDES WET program? If so, why, and if not, why not?

All of the commenters concurred that the bioequivalence method used in this study is a sound conceptual approach. Most also agreed that the TST approach is an improvement over the current accepted hypothesis testing approach used in the NPDES WET program. Commenters raised the issue of the method for selecting the value for *b*. Commenters also offered opinions on data analysis: a limitation of real world data is that estimated error rates are based on sample data means and not population means; without an objective standard of comparison, although a reasonable exercise, empirical studies can only provide a comparison of the methods.

All chronic WET methods now use a b=0.75, reflecting a risk management decision based on ecological knowledge and recent research. Data analysis now includes extensive simulation as well as use of empirical data.

EPA Question 2 - Document Responsiveness

Assess whether the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis.

Four of the five commenters agreed the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis. A dissenting commenter believed that hypothesis testing and TST approaches are not all that different in so far as that both approaches are based on experimental designs reflecting the magnitude of the effect sought.

Dissenting commenter is correct up to a point. As explained in the revised document, TST empowers the permittee to do more than the minimum test design if desired.

EPA Question 3 - Document Data Analysis Basis

Assess whether the data supporting the recommendations and conclusions on the draft TST *document are technically correct and defensible. The draft TST document attempts to evaluate existing data comprehensively, but: (1) for the purposes of standardizing comparisons, relies on data developed after 1995; (2) to be comprehensive, evaluates data developed using EPA WET test methods conducted under the current 2002 edition, as well as some earlier editions; and (3) to ensure that conclusions are based on appropriate data, censors some data points. The Agency's reasoning behind each of these aspects of the evaluation is explained in the draft document and related references (i.e., data test acceptance and quality assurance protocol).*

All of the commenters generally agreed that the data supporting the recommendations and conclusions are reasonable and defensible. There was a consensus that it was better to focus on current methods and future data. Commenters had an issue with an apparent assumption of normal distribution, and questioned if the assumption of a normal distribution can be made for all data used. A commenter had a set of specific issues critical of the documentation in the section of the document on the simulation method.

The peer review document as well as the revised document focus on data using current WET methods. Future data are no longer a consideration because *b* is now independent of test method performance. The revised document now discusses non-normal data and how these are handled using the TST approach. The simulation method section is described more clearly and in greater detail.

EPA Question 4 - Document Conclusions

Assess whether the draft TST approach as applied is technically defensible especially if challenged by either the NPDES regulated community, permitting authorities or expert consultants hired by permittees or other interested parties. Specifically, bioequivalency "b" values were derived for each test method using several risk management decision criteria which together, were intended to balance desired maximum alpha and beta errors at specific mean effect levels and within-test variability. Comment on the fact that this draft TST approach could be similarly used for additional WET test method(s) in the future. This draft TST approach builds upon EPA's earlier peer reviewed NPDES WET Variability document (USEPA 2000e) to derive and evaluate the "b" values. Evaluate the methodology used in the draft TST document to derive method-specific "b" values and apply the draft TST approach.

Commenters generally agreed that the TST approach is technically defensible. All commenters agreed, however, that the method in determining "b" values is critical to the validity of the TST approach and its acceptance by the regulated community. A commenter was concerned

specifically about the way "b" was determined and also about the alpha values chosen. A commenter suggested including different statistical distributions to improve the robustness of the simulation results. A commenter suggested dropping the label "bioequivalency" as it conveys a potentially confusing meaning due to its historical use that is unnecessary to its meaning or use in the TST approach.

All chronic WET methods use b=0.75 based on ecological knowledge and consistency with the current WET program (i.e., IC_{25}). Alphas are now chosen for broad groups of WET methods that achieve desirable error rates and fullfil risk management goals. Different statistical distributions (e.g., normal, binomial, etc.) have now been included in analyses. "Bioequivalency" has been retained in the revised document because TST now follows closely this approach.

EPA Question 5 - Document Quality Overall

Provide any recommendations for how this draft TST document should be presented to the public (or the users of this approach) particularly NPDES regulatory authorities such as NPDES States and EPA Regions (the document will be revised to accommodate readers with a more Plain English version). Suggest, if possible, how it's highly technical content should be translated into a version more readily understood by the NPDES regulatory public (again meeting EPA's Plain English requirements) and yet maintain its clarity given its potential scientific, regulatory, and technical applications. Also critique whether a regulatory authority and their permittees would clearly understand the draft TST document's recommendations and if not how specifically should it be revised to make it easier to implement under EPA's NPDES permit's program.

Commenters were generally sensitive to how the document should be presented to the public. They also were quite critical of the presentation and clarity of the draft document. All of the commenters had substantial issues with the clarity, completeness, and grammatical errors that they found to be common throughout the document. One commenter noted the document needed to be re-written before a "plain English" assessment was attempted.

The entire document has been significantly revised for content, presentation, clarity, completeness, and grammar. A shorter, less technical companion document is being prepared for regions, states, and the public.

EPA Question 6 - Recommendations

Provide any recommendations to improve the draft TST document's technical basis and approach for deriving the alternative WET statistical analysis method in the NPDES permitting program.

Commenters stated that many of their recommendations were presented in response to previous questions. Recommendations that were reiterated include the following:

- Eliminate the term "bioequivalence" because it will be met with resistance among NPDES permittees.

  We are using the bioequivalency approach, therefore, the term "bioequivalency" will remain. However, new text has been provided that presents TST and bioequivalency in a way that will not confuse.

- Present the decision criteria for selection of "b" in an explicit tabular form

All chronic methods now use b=0.75 so no need for decision criteria as before.

- Commit to monitor, analyze, and assess WET precision to refine alpha and beta error rates and "b" values

  All chronic methods now use b=0.75 and alpha (and resulting beta) is now clearly indicated in the revised document. There is now less need to closely monitor and assess on-going precision.

- Selection of the level of "b" should be by consensus

  b is now determined as an EPA risk management decision consistent with the current WET program.

- Present the TST method so that it is mathematically clear, including assumptions, steps, statistics, and criteria for evaluation

  Text has been re-written and presents the mathematics of the method much more clearly.

- Base simulation results on various non-normal distributions not just normal ones

  This has now been included in the revised document.

- Use weighted calculations to estimate "b" and power, not simulations

  b is no longer dependent on test performance including power.

- Cutpoints for "toxic" and "nontoxic" types of assessments are natural in the context of receiver operating characteristic (ROC) curves, and this type of analysis should have been included as part of this assessment.

  ROC curves were investigated and some aspects have been used in the revised document.

Response to Comments by Section

Executive Summary/Glossary

Commenter 3

Figure E-1 What is the purpose of the colors? If these mean something why is it not described in the figure legend?

Figure revised.

Page iv top low error rates than t-test should be lower error rates than the t-test.

Changed to "lower".

Figure E-2 What is the standard t-test. The two sample t with equal variance? What is meant by TST fails? It is difficult to see that the TST test is better unless a higher fail rate is better. In fact, one might argue that the degree of concordance is quite high for the two approaches.

Figure deleted.

The results of the project listed on the bottom of page iv really don't have anything to do with the project; these statements basically can be made just from reading papers on bioequivalence.

Context revised.

Table E-4. What is the difference between specificity and relative specificity? Here is the medical definition of relative specificity: The specificity of a medical screening test as determined by comparison with an established test of the same type. I am not even sure that what you have calculated is correctly termed specificity since specificity implies knowing the true state.

Table revised; we no longer use the term "relative" (in this context).

Page xi Glossary: why not use standard definitions of terms rather than creating misleading ones. For example: hypothesis test, power and significant difference (why connect this to a confidence interval) are not consistent with statistical definitions of these terms. Type I and II error are expressed in terms of hypotheses but power is not.

All definitions were taken from previous published EPA documents.

Commenter 4

Figure E-1, page iv: The labels for the symbols may be confusing to the audience. The phrase "NOEC passes," for example, implies that the NOEC test gave the correct answer, but really seems to be stating that the sample was categorized as non-toxic based on the NOEC test.

Labels have been reworded.

Figure E-2, page vi: The previous comment applies to this figure as well. Additionally, unlike the prior graph, it is unclear how the summarized results relate to the target percent effect. Perhaps this graph could be presented different effect level ranges, or effect level ranges could be included in footnotes under the graph.

Pie chart deleted.

Table E-2: Because the meaning of α and β differ between the two test approaches, it would be helpful to define them in a footnote for this table.

Table has been changed to be cleaner.

Table E-4, page ix: Footnotes 3 and 4 quote mean effect cutoffs of 20% and 25%, respectively. These should probably be the same.

Table has been revised. Footnotes no longer apply and are no longer needed.

Glossary, page xi: The definition of confidence interval is rather vague, though it may suffice for this document (as confidence intervals generally don't play a role in either the NOEC or TST approaches).

Definitions taken from previous published EPA documents.

Commenter 5

[title] The title of this report is somewhat misleading. The "test of significant toxicity" (TST) is compared to the "hypothesis testing' (HT) approach but not really compared to the "point estimate" approach. In addition, this newer approach is more of a test of equivalent toxicity (TET)

versus a test of significant toxicity (TST) if conceptualized from a bioequivalence perspective. The term "significant" can refer to either statistically detectable differences or to biologically meaningful changes.

TET was considered as an alternative title but TST was identified as a better title.

[ii] TST is simply known as "bioequivalence" in the literature. It is confusing and misleading to introduce new terms when this is already well described in the literature. Thus, the TST references should be changed throughout the report to bioequivalence.

"TST" relies on bioequivalence testing but refines it to accommodate the WET program.

[ii] HT is not equivalent to NOEC. Hypothesis testing is a general strategy for evaluating competing hypotheses and TST clearly employs HT as well.

Terminology edited.

[ii] The "advantage" of hypothesis testing may be a reflection of a common misinterpretation of NOECs and no-effect levels. The concentration associated with a response that is not statistically different from controls is not necessarily a safe concentration. The testing reflects the variability in the system, size of the effect that would be declared different, the power of the test, and the false positive/Type I error rate.

Text revised.

[ii] The so-called point estimate method (again an unfortunate label since confidence intervals are often constructed) yields what is more commonly considered a potency endpoint in other toxicology applications. Potency is estimated at a particular risk management (RM) level (e.g. IC25 or IC50) and there is an incentive to generate higher quality data since the CI for this endpoint would be narrower and the standard error for the endpoint would be decreased as well. The focus on the two-concentration test data may suggest the reason why this analysis alternative was ignored.

Deleted the point estimate method from the revised document.

[ii] The objective of finding the "b" for the TST to compare this to the HT approach implies that the "point estimation" approaches are not even in the mix any longer as is explicitly stated in the charge for this review.

This analysis does not compare TST to point estimate approaches. This is made clear in the revised document.

[ii] The Table E-2 summary is confusing. A figure or table might help with this description. For example, would something like the following help?

$0.75\times\mu_0$ [25% mean effect]	$0.8\times\mu_0$ [20% mean effect]	$0.9\times\mu_0$ [10% mean effect]	μ_0
$\mu_E<0.75\times\mu_0$	$\mu_E>0.8\times\mu_0$	$\mu_E>0.9\times\mu_0$	Mean response in reference / control group
Correctly Declare TOXIC 100%	Incorrectly Declare TOXIC < 5%	Incorrectly Declare Non-TOXIC < 20%	

There is a potential confusion between a true (yet unknowable) difference in population mean responses versus an observed difference in sample mean responses. This distinction may be lost on the reader. This is an important point since this exercise is based on observed differences in sample mean responses relative to observed variability.

This table and others like it have been deleted.

[ii] "β error rate" – this does not make sense. The errors that can be made in a decision framework are to reject the null hypothesis when the null is true (a Type I error or a False Positive error) or to accept null hypothesis when the alternative is true (a Type II error or False Negative error). The standard notation for the probability of these errors is α = Pr(Type I error) and β = Pr(Type II error). While most readers would be able to figure out what is meant, this type of statement is misleading and demonstrates insufficient care in presenting technical information.

Phrase removed from revised text.

[ii] It may be easier and make more sense to talk about this in terms of increased power vs. decreased Type II error rates.

Wording is clearer in revised document.

[iii] Reference to the simulation method here is a surprise. Early on in the summary, the focus was on empirical comparisons while here we see a simulation component mentioned.

The executive summary has been revised.

[iii] statements such as "TST never declared a 10% mean effect ..." should be restated as "TST never declared an OBSERVED 10% difference in sample means between effluent and control conditions ..."

The language has been revised.

[iii] Care must be taken when considering all of the percentage quantities described here. There are 1) observed sensitivity / power (%); 2) observed false positive/type I error rates (%); 3) the observed decrement in sample mean responses (% mean effect); 4) the value of "b" which is expressed as a % of the mean response; and related, the percentile of the CV distribution. This should be presented in a table or text box to make sure this doesn't lead to additional confusion.

Terminology has been made clearer and more precise.

[iii] Sensitivity and specificity are described without formally defining them. Unless the readers are familiar with health screening studies, these terms may be unfamiliar.

They have been added to the glossary.

[iii] The declaration of when a test was non-toxic appeared to be based on a subjective decrement from control/reference group responses.

This has been more clearly presented in the revised document.

[iv] If you are going to compare two methods for detecting toxicity in terms of error rates in the decision, then **receiver-operating-characteristic (ROC) curves** are the most common and natural way to display comparisons. In fact, the area under an ROC curve is a measure of the quality of screening procedure.

ROC curves were investigated and some aspects have been used in the revised document.

[iv] on the figure … "Percent effect in effluent" = $\mu_E/\ \mu_0 \times 100\%$ or = $(\ \mu_0 - \ \mu_E)/\ \mu_0 \times 100\%$ (assuming a continuous response characterized by a population mean, E=effluent 0=control and a decrease in response is adverse).

Figure edited.

[iv] on the figure … "CV Percentile" – refer to CV in control condition or the effluent condition? Are you assuming that these are the same?

Figure edited.

[iv] now describing α and β error rates vs. Type I and Type II error rates.

Error rates are described much more precisely.

[v] "rank a sample as toxic" – No, you are not ranking anything here. You are making a decision to declare a sample as toxic or not.

Language edited to reflect this comment.

[vi] These graphical displays are inappropriate and poor displays of the comparison of the two methods. Three-dimensional pie graphs are often criticized in the statistical graphics community (chart junk – more picture than data presentation – displaying a non-existent third dimension – etc.). More importantly, a table would be a much cleaner display here. For example, the first chart could be replaced by

		t-test (HT procedure)	
		Pass	Fail
TST	Pass	73.6%	3.2%
	Fail	4.9%	18.3%

This table provides a much better sense of concordance between the test results. In fact, we can easily see that the tests have a concordance of almost 92% here (concordance is a formal characteristic that is commonly defined in categorical data).

Figure deleted.

[vii] Table E-2. Mixing Type I and II error descriptors in the same column is confusing at best. These are observed rejection rates based upon various choices of "b." Extensive clarification is needed here.

Table has been revised.

[viii] Where is the Monte Carlo simulation analysis described? (Answer: Section 2.5.2) This was alluded to in the summary but not presented in this table.

Table has been revised.

[ix] Table E-4. What is relative specificity? Relative sensitivity? It is defined as "The specificity of a medical screening test as determined by comparison with an established test of the same type" by the American Heritage Medical Dictionary. Is that what you mean?

Table was revised to say "specificity" and "sensitivity"; "relative" was a typographical error.

[x] Add "HT" to list of acronyms? Add "aka bioequivalence" to TST?

Acronym has been added.

[xi] Glossary. A number of the definitions included in the glossary are imprecise and somewhat misleading and occasionally incorrect.

Confidence interval = interval estimate of a population parameter (not "around a point estimate of a population"). CIs can be one-sided or two-sided but this is not a critical point here.

EC = parameter that corresponds to the concentration of a toxicant associated with a specified level of impact. If a statistical model is fit, then the EC is derived from an inversion of the statistical model, i.e. a function of the regression coefficients. An estimate of this EC can be obtained after fitting the regression model.

HT = refers to using statistical hypothesis testing to identify, which if any, concentration condition differs from control conditions.

Alternative definition? HT "hypothesis testing method" – Using NOEC derived from t-tests (2 groups) or anova with multiple comparisons (if >2 groups) to evaluate mean differences under discharge and control conditions.

MSD = magnitude of difference OF WHAT? In the responses in an effluent group relative to responses in a control group?

NSEC/LSEC = I am not convinced that it is helpful to add more acronyms to the collection already in use in this arena.

RP = ? ecologically determined?

Significant difference = means of two distributions of sampling results? This is unclear. It appears to be defining the CI of the difference between two population means to be the Sig. Diff. Shouldn't significance be a function of an ecologically relevant change?

Type I Error (alpha)

Type II Error (beta) = alpha/beta are the PROBABILITIES of these errors, not the errors themselves.

All of the Glossary presentation appears to emphasize measured responses (continuous variates) versus proportions.

Definitions are from previous EPA documents.

Introduction

Commenter 2

p 7, l 3. How is statistical power incorporated into the TST? The statement is confusing because the TST also has a power. Which power is being referred to here? The TST and NOEC test have very different null hypotheses. The only connection (and hence the only way the power of the "usual" test of no difference is incorporated into the TST) is because all tests and all powers depend on the coefficient of variation. This claim is repeated many times in the document.

Statistical analyses have been re-done and this issue is addressed in the revised version.

Commenter 3

Table 1-1:

NOEC disadvantage 2: does not explicitly estimate statistical power – I don't think any method explicitly estimates power – I think you mean controls statistical power

Agreed, changed.

NOEC disadvantage 5 is also a problem with estimation of an IC50. It is also a potential problem with the equivalence approach.

Agreed.

Point estimate disadvantage 4: confidence intervals are also affected by assumptions

Point estimate columns deleted.

Page 1. Line 7: the NOEC is not a hypothesis test rather it is level (concentration or dilution) that is derived from a test of hypothesis.

Agreed, language changed to reflect comment.

Line 16: the LC50 is an estimated value and it seems to me answers the question at what concentration is the 50% effect predicted. If the WET test uses a criterion for evaluation that is based on a permitted IWC less than the IC25 it seems to be ignoring the uncertainty in the estimate of the IC25. Is this not a problem?

Not within the scope of this paper.

Page 4 line 9. Power depends on the variability of both groups (see figure 1-1). I think what you really want to describe here is the potential for a small biological effect being significant. If any decrease in survival relative to the control is an indication of toxic then is this not relevant?

Addressed in the revised text.

Figure 1.1 Very small intra-test variability – This would be better if there is not a pattern in the second group of data i.e. two lines of data. I would change the figure to show no linear pattern. The use of "very small" to characterize variability seems odd – why not just use "low".

No change was made in figure; it is only meant as an example. "very small" changed to "low".

Table 1-2B: I think "b" should be defined in the table since you take the opportunity here to define Type I and II errors. Should b<1 also be added?

I think the order of table and figures needs to be looked at as it seems that table 1-2B should come before figure 1.2

Figure 1.2 was deleted.

Page 6: It seems that the main value of the bioequivalence test is that it puts the burden on industry to use a sufficient sample size that has good power for the test. If sample sizes are not sufficient the effluent will not be declared "not toxic". Why would the approach be better than to require a post-hoc power analysis i.e. require the industry to have power of 0.8 for a change of say 30%?

True, they could. However, EPA has decided to use an alternate hypothesis approach instead. It should be noted that this is an optional approach.

Page 7: I am not sure what is meant by "maximum" desired alpha and beta rates. It seems one would want to minimize these or to be as close as possible to specified rates.

Sentence removed; no mention of "maximum desired α and β rates".

Page 8: the project objectives are not well written. The first paragraph is an awkward read. First, the primary objective (purpose) is stated (which seems to actually be two objectives). Then the second sentence lists another set of primary objectives.

1st paragraph rewritten.

Commenter 4

Figure 1-1, page 5: This is a useful graph for portraying the difference between the approaches.

Figure kept.

Commenter 5

[1] NOEC endpoint IS DEFINED BY A STATISTICAL hypothesis test that ... - The endpoint is not the HT approach.

Edited to read "NOEC endpoint is determined using a hypothesis test...".

[2] I don't agree with much of what is contained in this summary. Since the "point estimate" approach is not within the scope of the comparison, it is somewhat odd to see this included in the table. The point estimate approach alluded to here appears to be the ICp method and many of the criticisms relate to this method. The choice of effect level is no different than the choice of "b" in the TST as it is associated with some risk management level. The endpoint can be concentration dependent is listed as a disadvantage. I don't understand this criticism. Do you mean the spacing of concentrations may influence this? There have been 12+ years of scientific contributions to methods for aquatic toxicity testing that have appeared after this cited Pellston workshop, and these appear to be completely ignored in this report. One huge disadvantage of the HT approach is that people often misinterpret a NOEC as a threshold of no concern instead of an artifact of detectable effect sizes.

Point estimate columns deleted.

[3] MSD relates to Fisher's LSD or Tukey's HSD from multiple comparisons methods. It is the mean difference required to declare two population means different. This already includes the standard error of the difference in sample means. To further divide this by control mean is an attempt to give this a CV kind of interpretation.

MSD no longer used in analyses.

[4] Note that well designed experiments balance the Type I/II error rates. Again, these are NOT alpha errors and beta errors.

Language changed to be more accurate.

[5] The figures are a nice way to communicate that the same mean difference may not be declared different if the data are more variable. Although in both plots it is important to note that the effluent is DECLARED toxic or non-toxic. You don't know truth. You only know the outcome of this decision.

Agreed. Text edited to: "effluent is determined to be toxic" (or non-toxic)

[6] The null hypothesis is a statement about parameters of the population being equal and NOT an assertion that they are "not statistically significant." This is a fundamental concept and this type of mistake is fatal for this report.

Sentence edited from "not statistically significant" to "not different".

[6] What does "Treatment > Control" here mean? Are you only interested in one-sided alternatives?

Text revised to make clear.

[6] Table 1-2A. The footnote in this table is one of the few places where the error rates were carefully and correctly defined.

Agreed.

[6] Table 1-2B. "Effluent ≤ b*Control" is not precise or clear. Do you mean "$\mu_E \leq b^* \mu_0$"?

Equation changed to "Treatment mean ≤ b*Control mean"

[7] Sensitivity=statistical power? This first sentence is confusing.

Sentence deleted.

[7] I'm not sure how this picture clarifies the story about HT vs. TST. Note that the HT approach is sometimes referred to as a t-test approach and here as the NOEC approach. This type of switching of description will confuse the general readership of such a report.

Figure was deleted.

[8] As I commented earlier, the "b" factor should reflect important biological / ecological shifts in the population response.

Each test used b=0.75 as noted in response to previous comments.

[8] What does "degree of protectiveness" in objective 2 mean?

Defined in objective.

[8] If you can't compare TST to the "point estimate" approach, then how was it possible to have permits written using either HT or point estimate approaches?

Text concerning "point estimate" has been deleted.

Data Characteristics

Commenter 2

p 10, figure 2.1 Labeling of 'average' on the figures. If the dots are the 90th percentile, I can't see how the lines are the "average CV", as claimed. Perhaps these are the average 90th percentile, but that's not the average coefficient of variation.

Edited to say "average 90th percentile CV"

p 16, table 2-3. This is a repetition of table 1-2b.

Talbe 2-3 was deleted.

p 16, l -4. "Monte Carlo" is an adjective not a noun, unless you're referring to the city state.

Changed to "Monte Carlo simulation"

p 17, l 4. "ground truth" Is the truth known in the actual WET data? I doubt it. Without knowing the truth, how does the analysis of actual data sets "ground truth" the simulation?

Language revised, phrase deleted.

p 18, l 6. The preceding paragraphs describe the choice of effect size and variability. Properties of a statistical test also depend on the degrees of freedom for the error (i.e. whether variances are pooled, the number of treatments, and the number of replicates). These values affect the power and the relationship between the TST and the usual test. Table 2-1 gives the **minimum** numbers. This makes me suspect that there was variability among tests. How did you choose the rest of the details for each test? If your calculations used a single degree of freedom error, say so, and report that degree of freedom error.

Explained better in the revised document.

p 16, lines 12 – end of page. This is really important stuff because it describes how you translated risk management criteria into an evaluation of both tests. It is a key feature of your analysis. It must not be buried inside a section described as 'Freshwater and East Coast ...'. I found this section very hard to understand. Since it lies at the heart of your evaluation, it needs to be prominent and clearly written.

Agreed. The section was rewritten and reorganized to make more visible.

p 19, l 4. This sounds like a fourth criterion. (or are you implying that a test should satisfy all criteria). If so, say that.

Criteria were completely changed in revised document.

p 20, l 8. Where is exhibit A?

Exhibit A is an example, which has been made more prominent in the revised document.

p 20, l 15. The arc-sine square root transformation is a **variance stabilizing** transformation. It does not correct for non-normality.

Agreed, text changed to reflect comment.

p 20, l -10. I don't see why four replicates matter here.

Agreed. We have revised the analysis to evaluate two as well as four replicates.

p 20, l -8. You simulated data for four replicates, even though reality is only 2 replicates. That means you are reporting properties for a non-existent statistical test (using 4 reps).

Agreed, revised the analysis.

p 20, l -3. This seems like you're trying to force non-normal data into a t-test framework. This introduces all sorts of complications, as you discuss. However, I'm not sure that alternatives (e.g.

binomial exact tests or beta-binomial tests) are any easier. They certainly are not part of the standard statistical toolbox.

This issue is further discussed in the revised document.

p 21, l 8. I'm not sure what you're doing here. In fact, I'm completely confused by what you say you're trying to do. The problem with establishing alpha and beta from actual data is that both require that you know the true difference (is it zero or not). You don't know that.

One ways to bypass this problem is to use only the control data and add a specific effect into the effluent mean. I couldn't tell whether you used observed test data 'as is' (without knowing truth) or whether you added known values to control data.

Analysis approach was revised and this is no longer an issue.

p 21, l -6. This is a very unusual definition of alpha level. It isn't clear whether 'exceeding the toxicity threshold' applies to population quantities or sample means. If this is based on sample means, the computation is completely wrong, since sample means often exceed (and even as likely as 50% to exceed) the population quantities.

Wording was revised.

Also, aren't the definitions backwards, since % effect is calculated as (control – effluent)/control, so a large number is a 'bad', i.e. toxic result.

Agreed, re-worded.

p 21, l -2. The number of data sets has nothing to do with their appropriateness for evaluating the simulations.

Re-worded.

Figure 2.2. This is one of many examples of poor graphics. Specific problems include:

a) the legend describes %, with a range of 0 – 100%, but the y axis is scaled from 0 – 1.

b) the y-axis label is the title of the plot, not a description of what is plotted on the y axis.

c) the two lines are redundant. One is 1- the other.

Figure deleted.

p 23-24. 3 issues:

a) Isn't this out of place and said earlier?

Text has been moved.

b) Distributions of t: these aren't very useful, since the t statistics are very different for the two tests.

Figure deleted.

c) l -2 on p 24. Where did 25% effect come from? b is 0.68, which translates into a 32% difference from the control. Same issue on next page.

Changed, all b values = 0.75 as noted in previous responses.

Commenter 3

Figure 2-1. Is it appropriate to talk about the coefficient of variation as test variability?

In reading the document, I was very confused by the connection between the "test" used to statistically evaluate an hypothesis, the data and the simulated data. I did not find in the document a formula for the bioequivalence test statistic. Is the test being used just a variation of the two sample t-test? If so, why not write it out. I think it also has to be connected to Table 2-1 which describes typical data for the study design. For an uninitiated reader it would be useful to know what the minimum number of effluent concentrations corresponds to. For the bioequivalence test, is the reference compared to all of these or just the full concentration? Table 2-1 needs to connect to section 2.5 and the test statistics. In addition, if different methods are applied to the data from these studies in different situations, should you not also consider different tests for bioequivalence?

Explained in more detail in the revised text.

The study is based on the assumption that the data represents a sample from laboratories or facilities that is in some sense probabilistic. This is needed to treat the screened sample (20 most recently conducted tests) as a sample from a population. How can one tell if this is an unbiased collection of observations? It might be useful to know what the universe of samples is and how the observed facilities and laboratories represent this universe. What are your assumptions?

Explained in more detail in the revised document.

Page 15: There is a need to be explicit as to the formula for the test statistics, the assumptions, degrees of freedom, etc., for the bioequivalence test.

The formulas are now explicitly stated.

Although beta=0.8 is probably the first size of power that comes to mind for most statisticians, there is also an argument for an alpha:beta compromise.

This has been reworked completely.

Table 2-3: I like that the hypotheses are defined in terms of parameters. However the mean of the treatment is defined incorrectly. It is not the response of the effluent concentration rather it is the mean response to or the population mean response to... Why then switch to effluent<b*control which is not very descriptive?

Table 2-3 was deleted.

Section 2.5.2 Is "several different" good grammar?

Was changed to only "different".

The notation used on page 23 bottom is confusing. Sc is referred to as the variance of the control yet in the formula Sc2 is used. Why use SDEVc rather than simply Sc. I would use the standard notation Sc2 for a sample variance and Sc for a standard deviation. Include degrees of freedom and critical value. I think it should be □=.05 not □<0.5. A clear way to write it is t.05(1),18=1.73. It also seems that Step 1 should be to state the hypotheses (this way one knows it is a one-sided test).

Example was redone.

I think the example using the unequal variance case (page 25) should provide a general formula not a formula for the special case when Nc=Ne. Otherwise this could be misleading for someone using the example as a template for their data.

This was changed.

Page 24: the standard statistical statement is "do not reject" the null hypothesis rather than "accept" the null hypothesis.

Agreed, changed.

Some things that are missing from the simulation example:

What is the sample size? Is the sample size for control treated the same as for the effluent?

Yes, and this is now made clear in revised document.

Was the control sample used with all levels of effluent mean and coefficient of variation (as in a block design) or was a new sample selected for each effluent level.

This issue is better explained in the revised document.

Was the choice of 1000 justified? For power calculations, I am used to having 10,000 simulations rather than 1000 as this sample size will lead to smaller simulation error.

Re-run to include 10,000 simulations

What are the formulas for the tests?

Formulas were inserted into the revised document.

What are the assumptions?

Assumptions are revised and clearly stated.

It appears that the mean and coefficient of variation were sampled independently. I think this is not justified. The mean and standard deviation are attributes of the individual WET test. I think they should be sampled together rather than separately.

The issue is discussed plainly in the revised document.

I also do not understand why one would sample mean and coefficient of variation rather than mean and standard deviation. For the test statistic, the standard deviation would have to be calculated. Are the mean and coefficient of variation independent?

The issue is discussed plainly in the revised document.

I question whether a Monte Carlo simulation is necessary. I think a more accurate approach is a direct calculation. Given the means, standard deviations, and a value of b compute the power of the test assuming normality. Then weight by the probability of the mean and standard deviation occurring in the population. Repeat this for a set of means and standard deviations (over a grid, say) and then compute the weighted average power. This approach should be more accurate.

Revised the simulation such that it addresses this issue.

It is also common to use cross validation in these cases to evaluate error rates. Since the value of b is calculated using all the data, shouldn't a test using the value of b be evaluated on a separate data set?

No longer relevant as all methods use b=0.75

Appendix A seems to have been pulled from the grant proposal as it is written in future tense rather than present or past tense. In reading this section, there is an implication that the EPA

flowcharts will be used to select a test. This may result for example, in a two-sample t-test assuming equal variance (which should have greater power that the test assuming unequal variance). It is not clear that the same approach is applied to the TST procedure. What is the justification for this approach?

Appendix A was deleted.

One additional criticism is that if the TST is an alternative to the WET approach, the simulation needs to compare the "approaches" not the "tests". The simulation as described compares the tests. The WET approach allows for a variety of tests depending on the assumptions. The conclusions therefore apply to a test not the approach.

Agreed, changes have been made accordingly.

Commenter 4

Figure 2-2, page 22: This figure (and all subsequent figures in this style) is a bit confusing. Depicting the TST failure rate and passing rate as two lines is redundant, because one rate is always 1 minus the other. Perhaps this information could be depicted as a set of stacked bar charts. Additionally, the legend states that the Y-axis is in percent units, while the axis label is in proportion units.

Figure deleted.

Section 2.5.1, page 16, 3rd sentence: "p" is not defined

Text revised, no longer refer to "p".

Commenter 5

[10] CV on the y-axis is for controls? Effluent group? Both?

Figure edited to be clearer.

[10] 90%-tile of CV from 1989 and 2000 – CVs from control group?

Figure edited.

[11] So what are you doing for the survival endpoints? What are you doing with counts? Are you using transformations (e.g. arc-sine-sqrt for proportions, sqrt for counts)?

This is better described in the revised document.

[12] Does this table imply that a control condition was run with each of these tests (e.g. effluent, reference toxicity)?

Yes. This is made clearer in the revised text.

[13] What does MSD mean for responses that are proportions (e.g. survival, germination)?

MSD was removed.

[15] Defining the mean response of the effluent here as μ_T should be done much earlier in this presentation. This would allow the bioequivalence/TST and HT approaches to be formally stated in terms of parameters.

Incorporated into the revised document.

[15] The level of change described as decision 3 is equivalent to the choice of "p" in ICp.

The use of "p" has been deleted.

[16] The description of the Monte Carlo simulation is inadequate and confusing. Monte Carlo was not used to simulate WET data. You used a Monte Carlo simulation to study the TST and HT approaches by first generating WET data with known underlying characteristics and then applying the approaches.

The description has been rewritten for clarity.

[17] second analysis EMPIRICALLY DERIVED Type I and Type II error rates ... with different "b" values defined for each calculation of these error rates.

No longer deriving different "b" values.

[17] Shouldn't you also simulate cases when the effluent mean was equal to the control mean?

Incorporated into the revised document.

[19] Type I error rate was equal to 0? [Here, incorrectly stated as α error=0.] This is an observed Type I error rate.

Terminology edited.

[20] It is bad statistical practice to talk about preceding a test of means with a test of variances. The F-test is notoriously sensitive to violations of the normality assumptions while the test of means are very robust. You can use an unequal variance t-test routinely as an alternative. Finally, other tests of variances such as Levene's test are preferred to the F test for variance homogeneity or Bartlett's >2 group generalization (assuming you want to formally test this which I believe is debatable).

Language changed to follow common statistical practice.

[20] 2 replicates vs. 4 replicates? What is a replicate here? Any reported simulation should be presented in sufficient detail so that someone could repeat your computer experiment. This presentation does not meet such a standard. Not only are the conditions unclearly presented but the implementation of the simulation is sketchy at best. For example, how was the simulation programmed (in Excel? FORTRAN? SAS? R?)?

Revised document is more detailed.

[21] mean effect levels versus an effect level defined as a change in mean response?

More carefully phrased in revised document.

[21] mean percent effect ranges? What are these? Why these ranges?

Better explained in the revised document.

[21] Not as robust statistically? What does this mean? You don't know "truth" in this empirical exercise. This comparison simply tells you how often the 2 methods lead to similar/dissimilar decisions.

Phrase removed.

[23] No, the test statistic is NOT formed from the population means μ_c and μ_e (what happened to the μ_T formulation earlier?) but in terms of sample means such as $\overline{Y}_C$

In this figure, doesn't nontoxic means $\mu_e > b\ \mu_c$

Example has been edited.

[23] The formula for calculating the pooled variance makes sense only if there are the same number of observations in both effluent and control groups.

Example has been edited.

[24] No, this is not the SE(mean) but the SE(difference in sample means)

Example has been edited.

[24] Doesn't the TST approach calculates the SE($\overline{Y}_e - b * \overline{Y}_C$)?

Example has been edited.

[25] No, the t-test statistics do NOT involve population means; they are functions of sample means. This is fundamental and critical notation.

Example has been edited.

[25] Doesn't b=0.68 imply toxic if $\mu_e \leq 0.68\ \mu_c$? Here, and elsewhere in the report, a decrease in response is considered adverse. Was this ever explicitly stated in the report?

Text has been edited.

[25] Isn't it better to say "equivalent to control response" instead of "not toxic?"

Agreed, example has been edited.

Quality Assurance

Commenter 2

p 26, all. This material needs to be combined with that in section 2.4, which also discusses QA.

Agreed, combined.

p 26. l 3. What is "Table 3", since tables are numbered as section-table?

Corrected in text.

p 26, l -8. What is computed in Excel?

Revised section completely.

Results

Commenter 2

p 29, table 4-2 (and all tables through 4-8 that are similar). These baffled me for far too long. The really important piece of information is the coefficient of variation for each test, which is hidden in footnotes. Nothing in this table made sense until I realized you were changing coefficient of variation for different rows.

I suggest a complete reorganization of this and comparable tables for other tests.

Effect level %	c.v.	Risk Criterion	B value				
			0.63	0.68	0.70	0.75	NOEC
15	≤ 75th%	Toxic < 0.2	0.0	0.001	0.003	**0.35**	**0.99**
20	25th-50th%	Toxic < ??	0.00	0.02	0.21	1.00	1.00
25	> 50th%	Non-toxic < 0.05	**0.99**	0.00	0.00	0.00	**0.78**
25	< 25th%	Toxic = 0	**0.90**	**0.19**	**0.62**	**1.00**	**1.00**
30	Any	Non-toxic = 0	**0.99**	0.00	0.00	0.00	**0.22**

This suggested revision a) includes the coefficient of variation as a specific element of the table, b) eliminates the meaningless columns (the –'s), and c) indicates where the risk management guidelines are exceeded (by the bold entries). Some of my entries for the 20% effect line are hypothetical, because I couldn't figure out the important information for that line in your table 4.2. If you feel that combining toxic and non-toxic endpoint s is too confusing, then separate the above table into a part for the toxic endpoints and another part for the non-toxic endpoints.

Table significantly revised.

p 30. Figure 4.1 Why didn't you use connected lines, as in figures 4.2 and 4.3?

Figure deleted.

Table 4.2 et seq. Is the value of 0.00 for the alpha level for 30% effect and b=0.70 correct? I think there is either a major failure in the computations or a major failure in the communication of how these results were computed. Here's why:

The usual interpretation of 30% effect is that control mean - effluent mean = 0.3 * control mean. When b = 0.70, then the effluent mean is exactly on the boundary of the equivalence region (because the boundary of the equivalence region is effluent mean = 0.7 control mean, i.e. control – effluent = 1 – 0.7 = 0.3 = 30% effect). When the population mean for the effluent lies exactly on the boundary of the equivalence region, the TST should have alpha = 5% no matter what the coefficient of variation is. This is follows from the construction of the test and the definition of the alpha level. I am very confused why the reported values are 0.00 (e.g. for 30% effect, b = 0.70, non-toxic (alpha) reported as 0.00).

Figures removed, only using b=0.75.

p 32, line -6. You refer to table 4-1. Shouldn't this be table 4-4?

Figures deleted.

Commenter 3

Page 27 why mention Ceriodaphnia is a freshwater invertebrate – water flea when this is mentioned in the header for 4.1. It is stated that 65% has power >= 0.8 and 35% is <=0.8. Should one of these be a strict inequality?

"freshwater invertebrate – water flea" was deleted. >0.8 was changed to a strict inequality

Commenter 4

Table 4-2, page 29: This table is unclear. It appears that the values under α and β correspond to the proportion of simulated analyses that were categorized as toxic (β) and non-toxic (α). When a number is presented, it implies that that result would be the "wrong" one given the simulated effect level, while a '-' implies that that conclusion was the "right" one, and therefore would not be an error. However, the 25% effect level includes proportions for both alpha and beta. From the footnotes, this appears to be due to the data being produced by different assumed CVs; however, this still implies that both results are "wrong." Much of this is likely due to the use of α and β, as they imply that the presented proportions represent "errors." It would be more meaningful to the audience to label the columns "test concludes toxic" and "test concludes non-toxic" without α and β, and explain that these two values will not add up to 1 in all cases.

Table was revised.

Figure 4-1, page 30: This figure is useful, but it may be helpful to add dashed lines at 0.95 and 0.2 to emphasize the target error rates. Also note that this figure does not include interpolation lines between points, while all subsequent graphs in this format do.

Interpolation and dashed lines added.

Commenter 5

[27] How was the MSD determined? How did you determine the power > 80%?

MSD was removed.

[27] A 1993 paper reported a similar result? Isn't this backwards and the 1996 paper reported a similar result to the 1993 paper?

"similar" changed to "comparable".

[29] Need to comment/formally define the relationship between sensitivity and Type II error rates? Between specificity and Type I error rates?

Defined in the revised document.

[31] The table legend needs to be enhanced here. For example, how is effect level defined here? What do the Risk Management columns mean here? Isn't "b" a RM decision?

Table significantly revised.

[32] How does "mean difference" in this figure relate to "effect size?"

Figure deleted.

Evaluation of the TST Approach for 2 Sample-Concentration Test Designs

Commenter 2

p 44. The pie charts are terrible graphics. The false 3D only hinders the visual interpretation. A table presents the information much more concisely. You essentially include the table information when you give the actual %'s for each category. A graphical alternative is a mosaic plot.

Chart deleted.

p 44. This approach, a pairwise comparison of TST and NOEC test results for each actual data set, would be a very good way to summarize results for all the 'actual data' analyses in the previous sections.

Agreed, incorporated into the revised document.

Commenter 4

Figure 5-2, page 45: This figure isn't as useful without knowing the observed mean effects for the two sets of data. While it can show NPDES permittees the benefits of reducing variability, for other segments of the audience what really matters is whether the approach gave the "correct" answer or not.

Figure revised.

Section 5.0, page 42, last sentence of 1st paragraph: The statement "the t-test, a type of hypothesis test, is not usually designed to minimize the rate of false negatives in the WET program" is a bit inaccurate. It should really say that studies for which t-tests are applied are not usually designed to minimize the rate of false negatives.

Revised wording.

Commenter 5

[45] A standard boxplot is a better display here (e.g. box with lines at Q1, median, Q3 and whiskers extending from min to Q1 and from Q3 to max).

Boxplot modified.

Conclusion and Recommendations

Commenter 2

p 47, lines 3 et seq. This seems to be based on the simulations. This claim would be much stronger if was based on the actual data. That is, you could use the actual WET data, compute both the NOEC and the TST and compare the results using the approach on p 44. More protective is then shown by t-test fails = 0 while TST fails = something larger than 0.

Agreed, incorporated into the revised document.

Figure 6.1 (and also E-1). Most of this graph is blank space because the largest % effect is 40%, but the Y axis maximum in 100%. Re-draw the graph with Y max = 40% to focus on the interesting stuff.

Figure was revised.

p 54, l -5. Again, I don't see how power is incorporated into the decision process because power of the t-test has little to do with the decision from the TST.

Explained further in the revised document.

Commenter 4

Table 6-1, page 50: These tables are the most useful for the public, as it gives the frequency of "wrong" answers under both approaches, though it would be more accurate to show the rates when the variability is larger as well. There may be some value in re-arranging the columns so the two fractions of samples incorrectly categorized as toxic are paired together, rather than the two rates for a given approach. However, this is not a vital change.

Table revised.

Table 6-2, page 51: It may be useful to include the sensitivity and specificity based on the NOEC approach for comparison purposes.

Table revised.

Commenter 5

[47] TST was a viable alternative prior to this empirical investigation. Bioequivalence has a long and well studied history with pharmaceutical applications.

Agreed.

Literature Cited

Commenter 2

p 55. l -5. Need year for the Grothe et al citation.

Added to citation.

Appendices

Commenter 3

Appendix B. Are results based on 1000 simulations?

Re-ran with 10,000 simulations.

The axis legend is often cutoff i.e. Minimum Significant Difference. The truncation of the labels occurs throughout the graphs in the appendix. Also the numbers on the y-axes on many of the graphs are difficult to see as the overlap the axis line.

Graphs deleted and replaced with new ones.

Commenter 4

Table B-1: Rather stating that the percents were calculated as the percent toxic using a separate result column, the table would be more readily understandable to state this in the column headings. The heading for this table also incorrectly states that the next page shows the percent of samples categorized as non-toxic.

Toxic column deleted. Heading corrected.

Table E-1: Why does this table show the percentage of tests categorized as non-toxic, while all previous ones show the percentage of tests categorized as toxic? This will confuse the audience.

Appendices have been removed–no longer relevant.

Appendix B, "Detailed Analysis" figures: The label says "TET" rather than "TST"

Figures revised.

Appendix Graphs: Many of the graphs appear to have been cutoff when inserted. For example, those on page 99 of the PDF.

Figures revised.